A Deep Learning-Based Time-Domain Approach for Non-Intrusive Speech Quality Assessment

Xupeng Jia and Dongmei Li
Tsinghua University, Beijing, China
E-mail: jxp12@mails.tsinghua.edu.cn, lidmei@tsinghua.edu.cn Tel/Fax: +86-10-62782693

***Abstract*—Objective speech quality assessment is an important component in speech processing systems. It can serve not only as an evaluation metric but also as a loss function in some deep learning-based systems. In this work, a novel deep learning-based non-intrusive speech quality assessment approach is proposed. Instead of using manually designed features or magnitude spectrum as input, the proposed method directly works on the time-domain waveform. The perceptual evaluation of speech quality (PESQ) is used as the learning target, and the network structure is designed referring to the PESQ calculation procedure. Multi-task training strategy is employed to optimize the network. Experimental results show that the proposed approach can yield high correlation to PESQ in both matched and unmatched situations. The proposed method can also be used as a non-intrusive estimation model for other speech quality or intelligibility assessment methods, such as the short time objective intelligibility (STOI).**

I. Introduction

Speech quality is a concept reflecting the listeners' feeling to the heard speech. There are two ways to measure the speech quality, which are subjective listening test and objective algorithms. The subjective listening test can get accurate results if it is carried out properly. However, it costs a lot of time and other resources. An objective algorithm is a flow of given calculation steps. It costs little time and leads to a result that has high correlation with the result of a subjective listening test. Thus the objective algorithms are widely used in the field of speech signal processing. The perceptual evaluation of speech quality (PESQ) [1] is one of the most famous objective algorithms. Despite of the good performance of PESQ algorithm, it has two shortcomings that limit its application. One is that it need a reference signal which is hard to acquire in some real-world scenarios. The other one is that the algorithm does not support gradient calculation. For example, to design a deep learning based speech enhancement system that directly optimize speech quality instead of mean square error (MSE) or signal-to-noise ratio (SNR), a speech quality assessment algorithm that supports gradient calculation is essential.

During the last twenty years, several non-intrusive speech quality assessment methods have been proposed [2-12]. Dubey and Kumar [2] extracted a set of auditory features including multi-resolution auditory model (MRAM) features, Mel-frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) feature, and trained a Gaussian mixture model to predict the speech quality score. Sharma et al. [3] also prepared a set of handcrafted features, and employed classification and regression trees (CART) to make the assessment. Although these non-intrusive methods have already achieved relatively good performances, the complex manually designed feature sets are unsuitable for the gradient calculation.

With the development of deep learning, it is believed that the neural network can extract effective feature representation by itself. Soni and Patil [5] used an auto-encoder to convert the spectrogram of a speech utterance into deep features, and trained a fully-connected network to estimate the quality score. Fu et al. [7] proposed the Quality-Net based on bidirectional long and short-term memory (BLSTM) network, which also took the spectrum as input. Lo et al. [12] replaced the BLSTM network with a structure that combined convolutional neural network (CNN) and BLSTM. These methods trained neural networks mapping the spectrum into the quality score and satisfied the requirement of gradient calculation.

Recently, time-domain speech signal processing achieved great success. Luo and Mesgarani [13] proposed a time-domain speech separation model named Conv-TasNet. It is an encoder-decoder based architecture with a temporal convolutional module inserted between the encoder and decoder. It improved the performance of speech separation by a large margin and surpassed the ideal time-frequency magnitude masking. Similarly, in the field of speech enhancement, Pandey and Wang [14] also proposed a time-domain speech enhancement model and achieved significant improvement of performance.

In this work, we propose a time-domain speech quality assessment approach. The main inspiration for this work is the success that time-domain deep learning methods have made in speech separation. This is reasonable because it avoids the loss of the phase information and gives the network a chance to learn a better feature representation. What's more, as the outputs of the time-domain speech separation models mentioned above are speech waveforms, a time-domain speech quality assessment model can be directly used as the loss function to further optimize these models.

The rest of this paper is organized as follows. In section II we describe the structure of the proposed time-domain model in detail. In section III we show some experimental results and make some discussions. At last in section IV we draw the conclusions.

II. Network structure

DNN-based speech quality assessment models have different network structures, such as multilayer perceptron (MLP)

[5], BLSTM [7] and CNN [12], et al. These models treat the neural network as a black box and the fitting capacity is a main factor influencing the performance. We believe that the combination of human inspiration and deep learning can lead to a better result. For example, Fu et al. took the advantage of the observation that the distribution of the frame-level quality scores varies with the utterance-level speech quality in Quality-Net [7]. They added a frame-level constraint to the loss function while the constraint weight was decided by the corresponding utterance-level score. In this work, we propose a network structure for speech quality assessment that refers to the calculation procedure of PESQ. We will describe the calculation procedure of PESQ briefly and then explain the network structure in detail.

A. Calculation procedure of PESQ

PESQ is a widely used objective speech quality assessment method. The calculation procedure can be described concisely as follows.

1) Time-domain preprocessing. This step includes the level and time alignment, as well as the filtering corresponding to human ear canal.
2) Auditory transform. After the time-domain preprocessing in step one, the signals are converted into time-frequency domain. This step includes the Bark spectrum generation, frequency equalization, equalization of gain variation and loudness mapping. After this step the signals are transformed to the perceived loudness in each time frequency cell.
3) Disturbance processing and cognitive modeling. The disturbance here means the absolute difference of the perceived loudness between the reference and degraded signal. This step includes discard of some long deletion sections, masking of the loudness and asymmetric processing.
4) Final score calculation. This includes calculating the error average over frequency and time, and using a simple formula to calculate the final score.

According to this procedure, we design a two-step network structure instead of a single network directly mapping the features into the quality scores. The first network tries to estimate the disturbance for each time frequency cell, and the second network converts the estimated disturbance into quality score.

B. Time-domain speech quality assessment

Although the speech quality assessment models using spectrum features in time frequency domain have achieved good performance, they are faced with the shortcoming that they lose the phase information. Considering that it has been proved that the phase distortion of the speech signal can impact the listeners' feeling [15], we believe that time-domain speech quality assessment without losing the phase information should result in a better performance.

The structure of the proposed neural network is shown in Fig. 1. The network consists of an encoder, a disturbance estimation module and a frame-score predicting module. The encoder is a 1-D convolutional layer. It encodes the input waveform with a frame length of 20 sampling points and a hop size of 10 sampling points. It includes 128 filters which results in a 128-channel output. Linear activation function is applied here. This encoder converts the waveform into a feature map which is similar to the spectrum. Thus the proposed neural network can deal with the time-domain waveform.

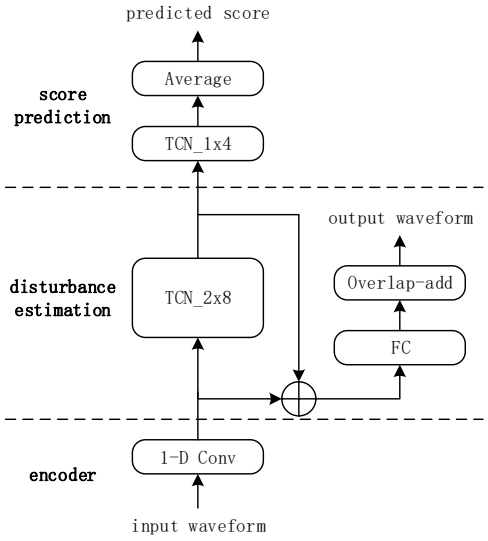

Fig. 1. Block diagram of the proposed neural network

The disturbance estimation module follows the structure of temporal convolutional network (TCN) which is used in [13] and [14]. It has two dilation blocks stacked together. Each dilation block consists of eight residual blocks with exponentially increasing dilation rate. The dilation rates are set to 1, 2, 4, 8, 16, 32, 64, 128 successively, resulting in a receptive field large enough for the disturbance estimation. The structure of the residual blocks is shown in Fig. 2. Separable convolutional layer [16] is used to reduce the model size, and another point-wise convolutional layer is applied before separable convolution to change the number of channels as well as compensating for the lack of fitting capability of the separable convolutional layer. The joint of the point-wise convolutional layer and the separable convolutional layer can reduce the number of parameters by 32% comparing to the normal convolutional layer, when the kernel size is set to 3. Layer normalization [17] is applied before every convolutional layer to optimize the training stage and restrain overfitting. Parametric ReLU [18] is used as the activation function. The output of the last residual block is fed into a fully-connected layer to predict the disturbance. Tanh is selected as the activation function of this layer.

The frame-score predicting module also follows the struc-

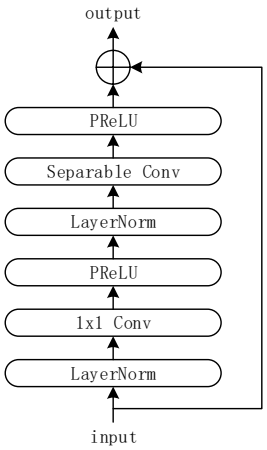

Fig. 2. Residual block used in the proposed neural network

ture of TCN, but with only 1 dilation block containing 4 residual blocks. The dilation block is followed by a fully-connected layer to predict the frame score. An average over all frames is calculated as the utterance score.

C. Multi-task training strategy

A multi-task training strategy is employed to optimize the proposed neural network. The main task is speech quality prediction and the assistant task is speech enhancement. The realization of the enhancement task is also shown in Fig. 1. The estimated disturbance is added to the output of the encoder to get an enhanced encoded feature. A decoder converts the enhanced feature into time-domain waveform. The decoder is simply realized by a fully-connected layer with linear activation function. At last the frame-level waveforms are overlapped and added together to get the utterance-level waveform. Based on this structure, the signal-to-noise ratio (SDR) is used as the loss function of the assistant task:

$$L_a = -10 * log(\frac{\|s\|^2}{\|\hat{s} - s\|^2}) \tag{1}$$

where $\hat{s}$ and s are the estimated and reference signal respectively, and $\|\cdot\|^2$ donates the signal power. The minus sign is added to the formula to let the optimizer minimize this loss function.

In previous deep-learning based speech quality assessment tasks, mean square error (MSE) is often used as the loss function. However, MSE and SDR are unmatched in scale, which may impact the training performance. We design a new loss function for the speech quality assessment task according to SDR:

$$L_m = -10 * log(\frac{c}{\|\hat{y} - y\|^2}) \tag{2}$$

where $\hat{y}$ and y are the estimated and reference speech quality score respectively, and c is a constant which is set to 1 in this work.

The overall loss function is the weighted sum of the main and assistant loss function:

$$L = \alpha * L_m + (1 - \alpha) * L_a \tag{3}$$

where α is the weight constant and is set to 0.5 in this work.

III. Experiments

A. Experimental setup

In our experiments, WSJ0 dataset [19] was used as the speech dataset. Two noise datasets were used. One is the NoiseX-92 [20] noise set and the other one is a self-made noise set named Noise-200. The Noise-200 dataset contains 200 different noise signals collected through the internet. It includes the natural sounds such as wind, stream, rain, bird, et al., and human sounds such as traffic, factory, cafeteria, hospital, cry, laugh, et al. The Noise-200 dataset was used to generate the train, valid and test sets, while the NoiseX-92 was only used in the test sets.

The train set included two subsets consisting of noisy and enhanced speeches respectively. The si_tr_s set in WSJ0 was correspondingly divided into two non-overlapping parts. 190 noise signals from Noise-200 dataset were used here. Note that the 190 noise signals did not include baby cry, rain, battle, cafeteria and factory noises, which were only used in the unmatched test set. For the noisy subset, the clean speeches were mixed with noise signals at randomly decided SNR levels between -10dB and 25dB. Each clean speech from the WSJ0 dataset was used only once. For the enhanced subset, after adding noise in the same way as the noisy part, a speech enhancement system was employed to suppress the added noise. The speech enhancement system has a similar structure with Conv-TasNet [13], but was modified to adapt to speech enhancement task. The speech enhancement system was trained on 12000 noisy speeches with 200 different noises at six SNR levels (from -10dB to 15dB with steps of 5dB). and achieved 10.13dB improvement on the speeches with babble noise at 0dB SNR. The noisy part and the enhanced part were mixed together as the train set. The total length of the train set was 24.9 hours. The valid set was generated in the same way with train set, but using the si_dt_05 set instead.

Three test sets were generated. All of them were generated from the WSJ0 test set. Each of the speeches was used twice for each test set, generating noisy and enhanced speech respectively. The first test set, named test-1, was generated with the 190 noises used in generation of train set. This is the matched test set. The second test set, named test-2, was generated with the other 10 noises which were not used in train set. The last test set, named test-3, was generated with 5 noises from NoiseX-92 dataset, including m109, machinegun, pink, white and Volvo. Both test-2 set and test-3 set were unmatched test set.

To evaluate the performance of the proposed network, the Spearman's rank correlation coefficient (SRCC), linear correlation coefficient (LCC) and root mean square error (RMSE) were calculated based on the predicted and true PESQ scores.

B. Comparison of loss functions

Despite of SDR, MSE can also be used as the loss function in speech enhancement tasks. Meanwhile, the loss function for speech quality assessment task has three choices, which are MSE, (2) and (4).

$$L_m = -10 * log(\frac{\|y\|^2}{\|\hat{y} - y\|^2}) \tag{4}$$

Thus there are different combinations of the loss functions. The comparison result on test-1 set is shown in Table I.

It can be seen that the combination of (2) and SDR as the loss function has the best performance. SDR is better than MSE for the enhancement task. Equation (2) and (4) have an advantage over MSE for the quality assessment task. We think that there are two reasons for this phenomenon. The first one is that the value of (2) and (4) is close to the value of SDR, thus it's easier for the optimization algorithm to balance the weights between the loss of the two parts. The second one is that the logarithm operation amplifies the loss value when it's very small, providing a good reference for the training of the network. The difference between (2) and (4) lies in the numerator. Equation (4) uses the true value as the numerator, giving the samples with better quality higher weights, while (2) uses a constant as the numerator, giving the samples equal weights despite of their quality scores. Equation (4) may work better in the field of speech enhancement (which is SDR actually), while as for the speech quality assessment task, (2) is a better choice. Note that (2) is equivalent to log-MSE, with two constants adjust its value. The following experimental results are based on the combination of (2) and SDR.

C. Experiments on different test sets

The experimental results of the three test sets are shown in Table II, Table III and Table IV. Three competitive algorithms are included. The first one uses BLSTM network to map the spectrum into quality scores. It has the same structure with Quality-Net [7], and is represented by Freq-BLSTM, because it works on frequency domain features and uses BLSTM as its main structure. The Second one uses CNN to map the spectrum into quality scores. The CNN is the same with the TCN used in the disturbance estimation module of our work. It is represented by Freq-CNN. The third one is a time-domain method. It uses the encoder and an TCN based predictor to map the waveform into quality scores. It is represented by Time-CNN.

It can be seen that the proposed approach has the best performance over the three test sets consistently. There are some interesting observations in the results. The comparison of the first two methods shows that CNN works better than BLSTM for this task. The comparison of the second and the third methods shows that it is easier for the network to converge to a better point in frequency domain than in time domain. We think that it is because we use PESQ as the training target, and the frequency analysis is similar to STFT, which is used to generate the spectrum. This gives the network a good start point. However, it can be seen that although the second method works well on the test-1 and test-2 set, it suffers a glaring performance degradation on the test-3 set, indicating the lack of generalization capability. Focusing on the RMSE results, it can be seen that time-domain methods have better generalization capability than frequency-domain methods.

Both test-2 and test-3 are unmatched test, with noise signals not appearing in train set. The reason that the test results have an obvious gap may lie in the distribution of the PESQ scores. As for the test-3 set, it is generated with five noise signals from NoiseX-92 dataset. Experimental results on small sets generated by single noise signal show that the networks tend to have a good performance on the set from m109 noise, and perform poorly on the set from machinegun noise and Volvo noise. Consistent with this, the distribution of PESQ scores of the m109 set is the most similar to the distribution of the train set, while the distributions of machinegun set and Volvo set differ from the distribution of the train set significantly. This explains why the proposed method has similar RMSE results on test-2 and test-3 set while the SRCC results have a large gap.

D. Experiments on STOI

The proposed network structure is designed for speech quality assessment. PESQ score is used as the training target instead of the mean opinion score (MOS) of subjective listening test because we don't have the MOS dataset at the moment. As a complemental experiment, we also test the proposed

TABLE I
COMPARISON OF DIFFERENT COMBINATIONS OF LOSS FUNCTIONS

Quality loss	Enhance loss	SRCC	LCC	RMSE
MSE	MSE	0.927	0.934	0.298
MSE	SDR	0.935	0.939	0.286
Equation (4)	SDR	0.939	0.946	0.269
Equation (2)	SDR	**0.942**	**0.948**	**0.264**

TABLE II
COMPARISON WITH OTHER METHODS IN RMSE

RMSE	test-1	test-2	test-3
Freq-BLSTM [5]	0.329	0.312	0.436
Freq-CNN	0.274	0.291	0.528
Time-CNN	0.318	0.324	0.345
Proposed	**0.264**	**0.284**	**0.277**

TABLE III
COMPARISON WITH OTHER METHODS IN SRCC

SRCC	test-1	test-2	test-3
Freq-BLSTM [5]	0.908	0.937	0.801
Freq-CNN	0.940	0.948	0.728
Time-CNN	0.913	0.928	0.866
Proposed	**0.942**	**0.952**	**0.917**

TABLE IV
COMPARISON WITH OTHER METHODS IN LCC

LCC	test-1	test-2	test-3
Freq-BLSTM [5]	0.920	0.931	0.854
Freq-CNN	0.945	0.940	0.803
Time-CNN	0.925	0.923	0.901
Proposed	**0.948**	**0.944**	**0.938**

TABLE V
COMPARISON WITH OTHER METHODS FOR STOI PREDICTION

	SRCC	LCC	RMSE
Freq-BLSTM [5]	0.786	0.867	0.061
Freq-CNN	0.832	0.909	0.051
Time-CNN	0.685	0.827	0.074
Proposed	**0.861**	**0.912**	**0.049**

structure on short time objective intelligibility (STOI) [21] estimation. The results are shown in Table V. It proves that the proposed structure also has advantages on the task of STOI predicting. It could be expected that the proposed structure would have a good performance on MOS predicting.

IV. CONCLUSIONS

This paper proposed a novel time-domain speech quality assessment method. Design of the structure followed the inspiration from the calculation procedure of PESQ, and multi-task training strategy was employed to optimize the network. Experimental results show that the proposed method has good convergence performance and significantly better generalization capability. It can also be applied to other tasks such as STOI prediction. Our future work includes training the propose network on MOS dataset and using it to improve our speech separation system.

REFERENCES

[1] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," International Conference on Acoustics, Speech, and Signal Processing, 2001: 749-752.

[2] R. K. Dubey and A. Kumar, "Non-intrusive Speech Quality Assessment Using Multi-Resolution Auditory Model Features for Degraded Narrowband Speech," IET Signal Processing, 2015, 9(9):638-646.

[3] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A Data-Driven Non-intrusive Measure of Speech Quality and Intelligibility," Speech Communication, 2016: 84-94.

[4] B. Patton, Y. Agiomyrgiannakis, M. Terry, et al., "AutoMOS: Learning a Non-intrusive Assessor of Naturalness-of-Speech," arXiv: Computation and Language, 2016.

[5] M. H. Soni and H. A. Patil, "Novel Deep Autoencoder Features for Non-intrusive Speech Quality Assessment," European Signal Processing Conference, 2016: 2315-2319.

[6] D. Yun, H. Lee, and S. H. Choi, "A Deep Learning-Based Approach to Non-Intrusive Objective Speech Intelligibility Estimation," IEICE Transactions on Information and Systems, 2018, 101(4): 1207-1208.

[7] S. Fu, Y. Tsao, H. Hwang, and H. Wang, "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM," Conference of the International Speech Communication Association, 2018: 1873-1877.

[8] X. Dong and D. S. Williamson, "A Classification-Aided Framework for Non-Intrusive Speech Quality Assessment," Workshop on Applications of Signal Processing to Audio and Acoustics, 2019: 100-104.

[9] R. A. Avila, H. Gamper, C. Reddy, et al., "Non-intrusive Speech Quality Assessment Using Neural Networks," International Conference on Acoustics, Speech, and Signal Processing, 2019: 631-635.

[10] B. Cauchi, K. Siedenburg, J. F. Santos, et al., "Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network," IEEE Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1151-1163.

[11] N. Parmar and R. K. Dubey, "Comparison of Performance of the Features of Speech Signal for Non-intrusive Speech Quality Assessment," International Conference on Signal Processing, 2015: 243-248.

[12] C. Lo, S. Fu, W. Huang, et al, "MOSNet: Deep Learning based Objective Assessment for Voice Conversion," arXiv: Sound, 2019.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," IEEE Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[14] A. Pandey and D. Wang, "TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain," International Conference on Acoustics, Speech, and Signal Processing, 2019: 6875-6879.

[15] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the Significance of Phase in the Short Term Fourier Spectrum for Speech Intelligibility," The Journal of the Acoustical Society of America, 2010, 127(3):1432-1439.

[16] A. Howard, M. Zhu, B. Chen, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv: Computer Vision and Pattern Recognition, 2017.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv: Machine Learning, 2016.

[18] K. He, X. Zhang, S. Ren and J.Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," International Conference on Computer Vision, 2015: 1026-1034.

[19] J. S. Garofolo, D. Graff, D. Paul and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Web Download, Philadelphia: Linguistic Data Consortium, 1993.

[20] A. Varga and H. J. M. Steeneken, "Assessment for Automatic Speech Recognition II: NOISEX-92: a Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," Speech Communication, 1993, 12(3): 247-251.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A Short-time Objective Intelligibility Measure for Time-frequency Weighted Noisy Speech," International Conference on Acoustics, Speech, and Signal Processing, 2010: 4214-4217.